Network Approach for Visualizing the Evolution of the Research of Cross-lingual Semantic Similarity

Aida Kh. Khakimova
aida_khatif@mail.ru
ANO «Scientific and Research Center for Information in Physics and Technique», Nizhny Novgorod, Russia

The paper is devoted to the problem of the bibliometric study of publications on the topic "Cross-lingual Semantic Similarity", available in the Dimensions database. Visualization of scientific networks showed fragmentation of research, limited interaction of organizations. Leading countries, leading organizations and authors are highlighted. Overlay visualization allowed us to assess the trends in citing authors. The expansion of the geography of research is shown. For international cooperation, the uniformity of semantic approaches to describing the concepts of critical infrastructure, incidents, resources and services related to their maintenance and protection is important. The stated approaches can be applied for visualization and modeling of technological development in the modern digital world. Semantic similarity is a longstanding problem in natural language processing (NLP). The semantic similarity between two words represents the semantic proximity (or semantic distance) between two words or concepts. This is an important problem in natural language processing, as it plays an important role in finding information, extracting information, text mining, web mining and many other applications.

***Keywords**: text mining, tech mining, cross-lingual semantic similarity, visualization, scientific network, bibliometrics*

1. Introduction

Linguistic similarities were studied by researchers from different fields using numerous statistical, linguistic and neuroscientific approaches.

The semantic properties of languages are usually evaluated using the embedding of words, which projects a linguistic dictionary onto the vector space of a given number of dimensions, in which the semantic relations of words are stored.

In artificial intelligence and cognitive science, semantic similarities were used for various scientific assessments and measurements, as well as for decoding complex interfaces of conceptualizing feelings [1].

Theoretically, semantic similarity refers to the idea of commonality in the characteristics between words or concepts in a language. Although this is a property of the relationship between concepts or feelings, it can also be defined as a measurement of the conceptual similarity between two words, sentences, paragraphs, documents, or even two parts of a text.

Recently, there has been a growing interest in finding semantically similar words in different languages based on comparable data easily accessible from the Internet (for example, Wikipedia, news) [2, 3].

According to Hotho et al. [4] Text Mining can be defined - like data mining - as the application of algorithms and methods from fields of machine learning and statistics in texts in order to search for useful templates after pre-processing. Data mining algorithms can be applied to the extracted data.

Text analysis in big data analytics is becoming a powerful tool for processing unstructured text data, analyze it to extract new knowledge and identify meaningful models and correlations hidden in the data. Text mining refers to the extraction of information and implicit patterns previously unknown in automatic or semi-automatic mode from a huge unstructured text data such as natural language texts [5].

Tech Mining refers to the application of text mining methods to technical documentation. For the purposes of patent analysis, this is called "patent mining". Tech Mining (TM) [6] uses text mining software to exploit scientific and technical information resources. Mining technology is used to inform technology management. This technology combines understanding of technological innovative processes with software tools for obtaining vital scientific and technical knowledge.

Whereas many applications have employed certain similarity functions to compute the semantic similarity between terms, most of the traditional approaches solving the problem by using dictionaries such as WordNet. The main problem is that a lot of terms (e.g. abbreviations, acronyms, brand names etc.) that are not covered by these kinds of dictionaries [7]. As a result, semantic similarity measures which are based on this type of resources cannot be used directly in these cases.

Tech Mining is the application of text mining tools to scientific and technical information resources. The ever-growing volume of scientific results represents a boom in technological innovation, but also complicates efforts to obtain useful and concise information for solving problems. This problem extends to technological mining, where the development of methods compatible with big data is an urgent problem.

In the current patent analysis, numerous patent documents use different words to describe the same event, leading to semantic inconsistency and polysemy due to the many meanings that may exist for a single word. To solve this problem, document analysis often requires combining synonyms into the same semantic dimension. On the other hand, different words can be used to describe the same events.

The methods for measuring the semantic similarity of texts are necessary for the development of areas of information retrieval, data mining and text analysis. Such methods will help to avoid patent infringement in the development of technological capabilities to achieve future competitive advantages [8].

The growing popularity of data science is also affecting high-tech industries. However, since they usually have different core competencies - the creation of cyberphysical systems, and not, for example, machine learning algorithms or data mining - to delve into the

science of data by specialists in the field, such as system engineers or architects, can be more cumbersome than expected.

In recent years, in order to help subject matter experts use data science, scientists have been developing semantic search engines. So, for example, Semantic Snake Charmer (SSC) [9], is a search engine based on subject knowledge. SSC includes a natural language processing module that can convert relevant documentation into several types of semantic graphs.

2. Related works

An accurate assessment of the actual similarity between documents is fundamental for many automatic text analysis applications, such as thesaurus generation [10], machine translation [11], question-answer [12], information search [13], and automatic generalization.

Semantic space is an attempt to model the characteristics of human semantic memory, which is guided by the principle that words with similar meanings are found in a similar language environment. Semantic space is a vector space that captures the value quantitatively from the point of view of coincidence statistics, where words (or concepts) are represented as vectors in a high-dimensional space [14]. As a result, the similarity of the meanings of words can be quantified by measuring their distance in a high-dimensional vector space.

Latent semantic analysis (LSA) is based on the fact that words that have similar meanings tend to occur in similar texts [15].

Knowledge-based methods suffer from a limited number of common vocabulary words that are commonly used in general English literature and often not suitable for specific domains.

The vector space model is classically used to evaluate the semantic similarity between two documents. Terms are represented in this semantic space as vectors called word embeddings. The possibilities of determining textual similarity based on vector representations of terms in a semantic space in which the proximity of vectors can be interpreted as semantic similarity [16] are investigated.

The LSA method has an advantage over most modern information retrieval methods because it has the ability to measure the similarity of two texts that use completely different words. However, there are morphological problems of the correct identification of terms, as well as more fundamental problems with homonymy / polysemy and synonymy. Techniques that depend on large enclosures tend to overestimate relatively unrelated sentences or relatively related sentences (e.g., LSAs). LSAs overestimate the similarity score of compared pairs of sentences [17]. The study of the similarity assessment between patent documents and scientific publications in the field of biotechnology by the LSA method proved that in this case the decrease in dimension led to the cutting off of valuable information [18].

Semantic spaces can be constructed either using the additive model or the multiplicative model. Both additive and multiplicative approaches to constructing semantic space do not take into account the word order among the components (i.e., words or phrases). Traditional clustering algorithms usually rely on the BOW (Bag of Words) approach, and the obvious drawback of BOW is that it ignores the semantic relationship between words.

Researchers expanded DSM to include the compositional structure of the language, and called these models compositional-DSM (CDSM). CDSM models suggest that the meaning of a word can be interpreted by its context, and the meaning of a sentence can be obtained from its compositions [19]. The central place in CDSM is compositionality, that is, the meaning of complex expressions is determined by the values of their component expressions and the rules for combining them.

Assessing semantic similarities between concepts is a key tool to improve understanding of texts. The structured knowledge provided by ontologies is widely used to evaluate similarities. However, in many areas several ontologies modeling the same concepts in different ways are available. The paper describes the criteria for choosing ontologies for assessing semantic similarity [20].

A measure of calculating the similarity between sentences or between documents using an ontology is proposed. The similarity is evaluated using the concept vector of the document (proposal), formed by finding the links between the ontology terms and the content of the document (proposal) [21].

The vector space model is used to identify potentially useful services and evaluate web services [22]. Methods for extracting information and automatic semantic textual similarity assessment were used for electronic health systems (EHR) [23].

Similarity measures are used to select a context-sensitive application that matches the current context of the user. Personalization of services is directly related to the user's preferences, displaying his contextual information from the user environment.

A semantic similarity measure is a tool for assessing the similarity between instances of the context, which allows to select services in accordance with their relevance for a given request, profile and user preferences. With this approach, the context is considered as a set of information representing spatio-temporal information about the user, as well as his preferences and interests, which is used as a factor in classifying services by relevance [24].

The data sets of common STS problems were widely used to study similarities at the sentence level and semantic representations [25-27].

The CL-WES method [28] is based on the cosine similarity of distributed representations of sentences, which are obtained by weighting the sum of each word vector in a sentence. At the same time, at the first stage, the Spanish sentence is translated into English using Google Translate (i.e., two sentences are formulated in the same language), then both statements are compared.

The similarity score of the interlanguage pairs in English and Spanish was calculated as the average of the corresponding language ratings in the monolingual data sets [29]. The study was developed for five languages [30] - English, German, Italian, Spanish and Farsi.

The skip-gram model has become one of the most popular for the study of word representations in NLP [31].

The cross-language definition of semantic textual similarity is an important step for the detection and evaluation of interlanguage plagiarism; research in this area is rare.

A comparable corpus consists of documents in two or more languages or varieties that are not translations of each other and deal with similar topics. Comparable bodies are, by definition, multilingual and interlanguage collections of text. The Internet can be used as a huge resource of multilingual texts.

3. Matherials and methods

To search for publications, the Dimensions database (https://app.dimensions.ai/) was used, which provides open access to more than 95 million publication records and related metrics for individual users. The search keywords used were “cross-lingual semantic similarity”. 2050 articles were discovered.

VOSviewer (https://www.vosviewer.com/) was used to visualize scientific networks. VOSviewer uses a remote approach to visualizing bibliometric networks. In a bibliometric network, there are often large differences between nodes in the number of edges they have to other nodes.

Popular sites, for example, representing highly cited publications or highly prolific researchers, may have several orders of magnitude more connections than their less popular counterparts. When analyzing bibliometric networks, normalization of these differences between nodes is usually performed. VOSviewer by default applies the normalization of communication strength [32].

4. Results and Discussion

2050 articles of 2825 authors from 64 countries were discovered. The dynamics of publications is shown in Fig. 1. The trend line is clearly exponential, the determination coefficient (R^2), which is also called the approximation confidence value, is 0.6648. Initial publications date back to the 80s of the 20th century, but research has been growing since the beginning of the 21st century.

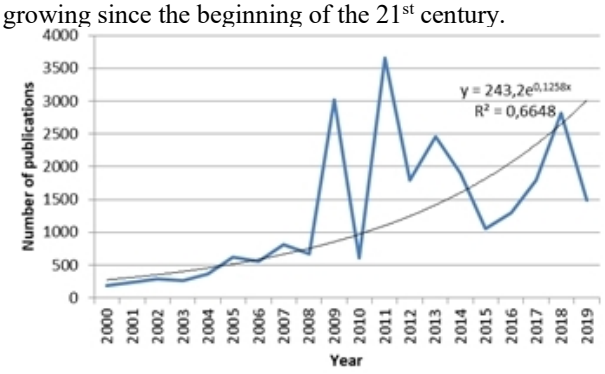

Fig. 1. Dynamics of the number of publications devoted to the problem of cross-lingual semantic similarity

With the help of VOSviewer, a co-authorship network was built. For 2825 authors, the minimum number of articles by the author was taken to be five; 26 such authors were identified in 17 clusters. The largest cluster included 4 authors. Fig. 2 shows that there is a separation of the authors into small research groups.

Fig. 2. Collaboration Network

We reviewed a collaborative network of organizations (Fig. 3). For 684 organizations, the minimum number of articles of the organization was taken to be five; such organizations were allocated 64 in 6 clusters. Fig. 3 shows that only a small number of universities interact. The largest cluster included 11 European universities and organizations: Dublin City University; Fondazione Bruno Kessler; German Research Center for Artificial Intelligence; National University of Distance Education; Trinity College Dublin; University of Alicante; University of Edinburgh; University of Sheffield; University of The Basque Country; University of Trento; University of Wolverhampton.

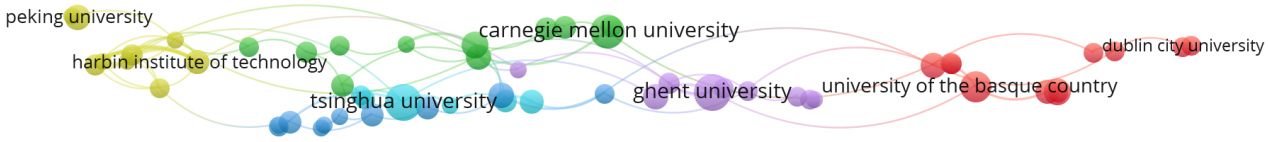

Fig. 3. Collaboration on organizations

We examined a co-authorship network by country, the minimum number of articles by the author was taken to be five. Of 2825 authors of 64 countries, 35 are associated in five clusters (Fig. 4). The two largest clusters included 9 countries. The first cluster included countries: Austria, Canada, China, India, Iran, Japan, the Netherlands, Taiwan, and the USA. The second cluster included countries: Belgium, Finland, France, Greece, Slovenia, Spain, Switzerland, Tunisia, Great Britain.

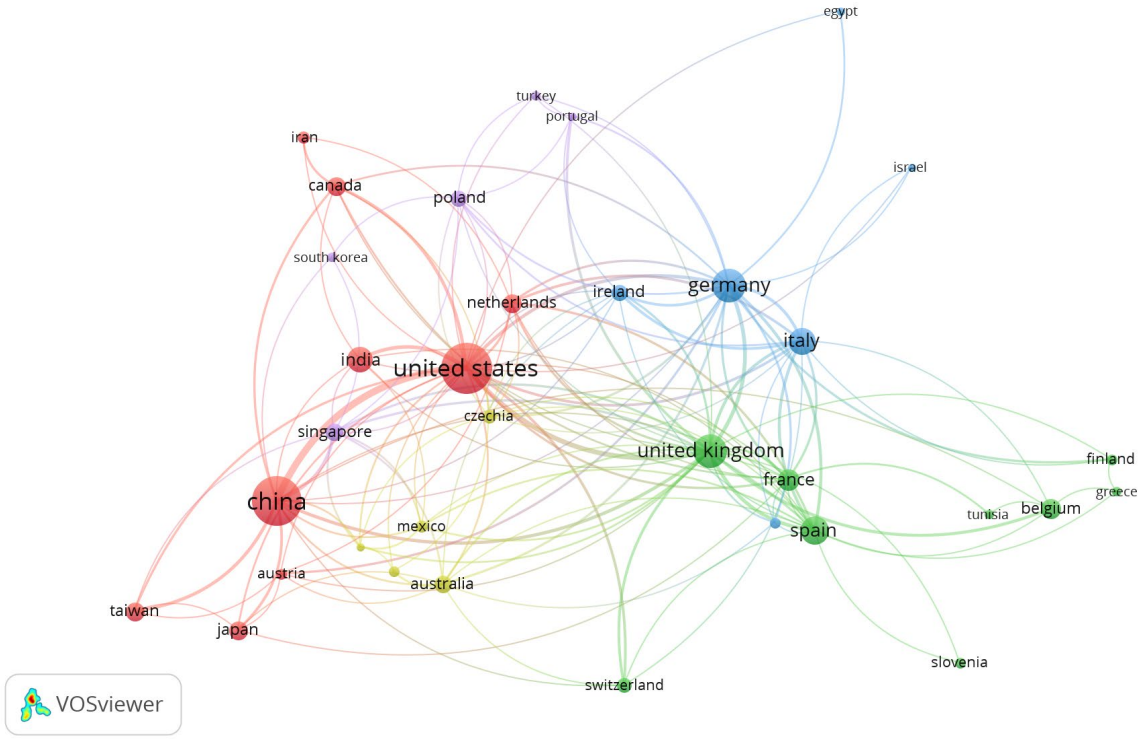

Fig. 4. Co-authorship by country

The citation index in recent years is the main measure of the value of both a scientist and an institution, so we examined citation networks.

We examined the citation network for documents, the minimum number of publications by the author was taken equal to ten. 298 authors from 2050 were identified in 14 clusters (Fig. 5).

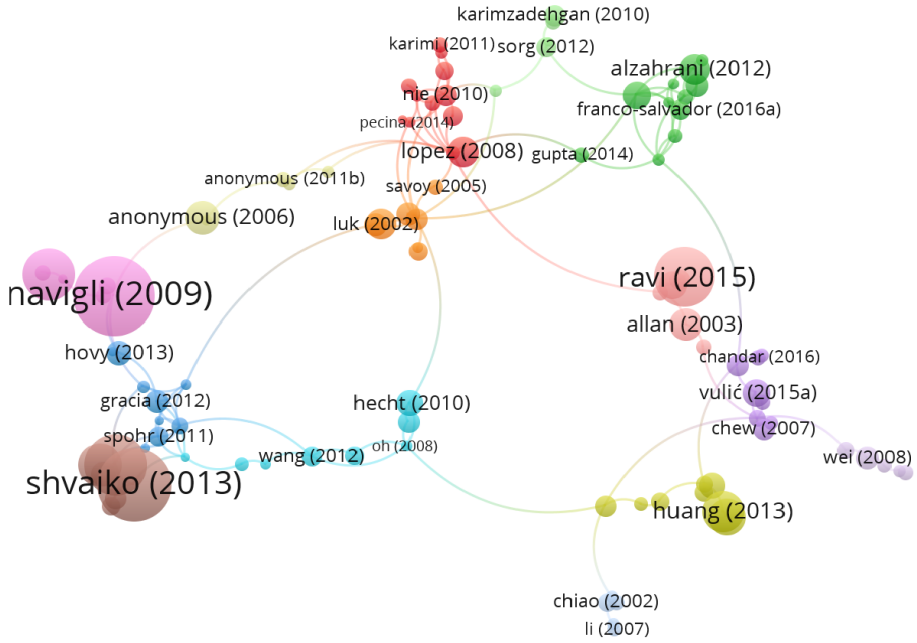

Fig. 5. Citation from publications of the most cited authors

The most cited author is Navigli, Roberto (759 citations) [29, 30]. More than 200 citations from Rosso, Paolo (239) and Moens, Marie-Francine (216) [2].

VOSviewer also supports overlay renderings. In overlay rendering, the color of a node indicates a specific property of the node, for example, the year of publication. We presented the authors citation network in an overlay visualization option to assess citation trends (Fig. 6). The figure clearly shows that R. Navigli is the founder in the area.

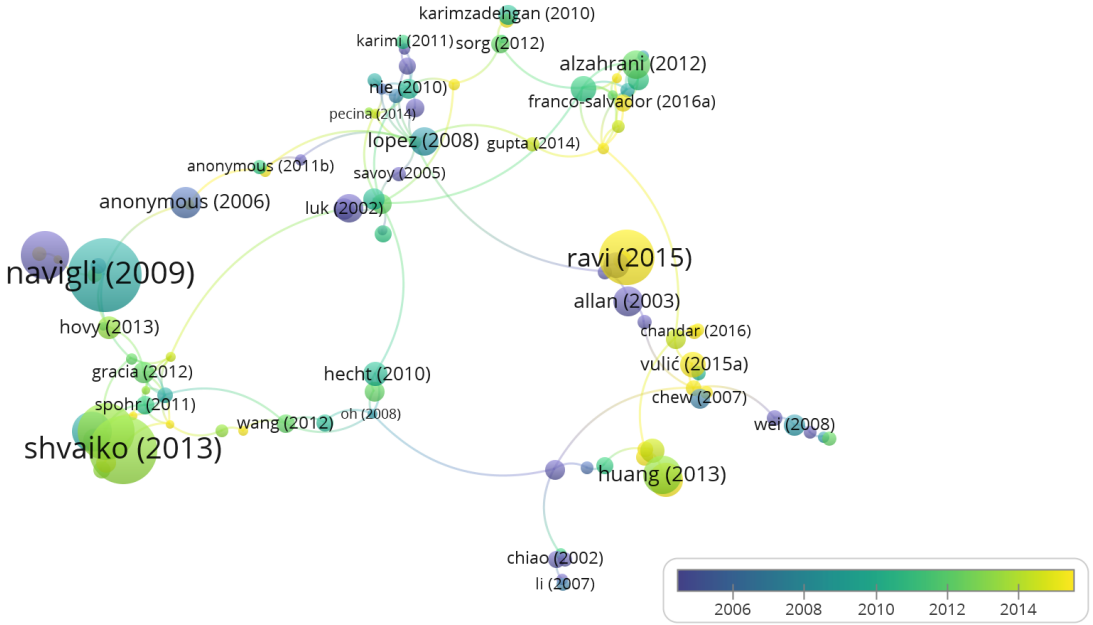

Fig. 6. Overlay citation network visualization

A citation network by authors was built. The minimum number of publications by the author was taken to be five. 16 authors were identified from 2050 in 2 clusters (Fig. 7). Mittal Namitali (2017), Rettinger Achim, Gipp Bela, Li Juanzi and Zhan Lei (2016) are the most recent of the most cited authors.

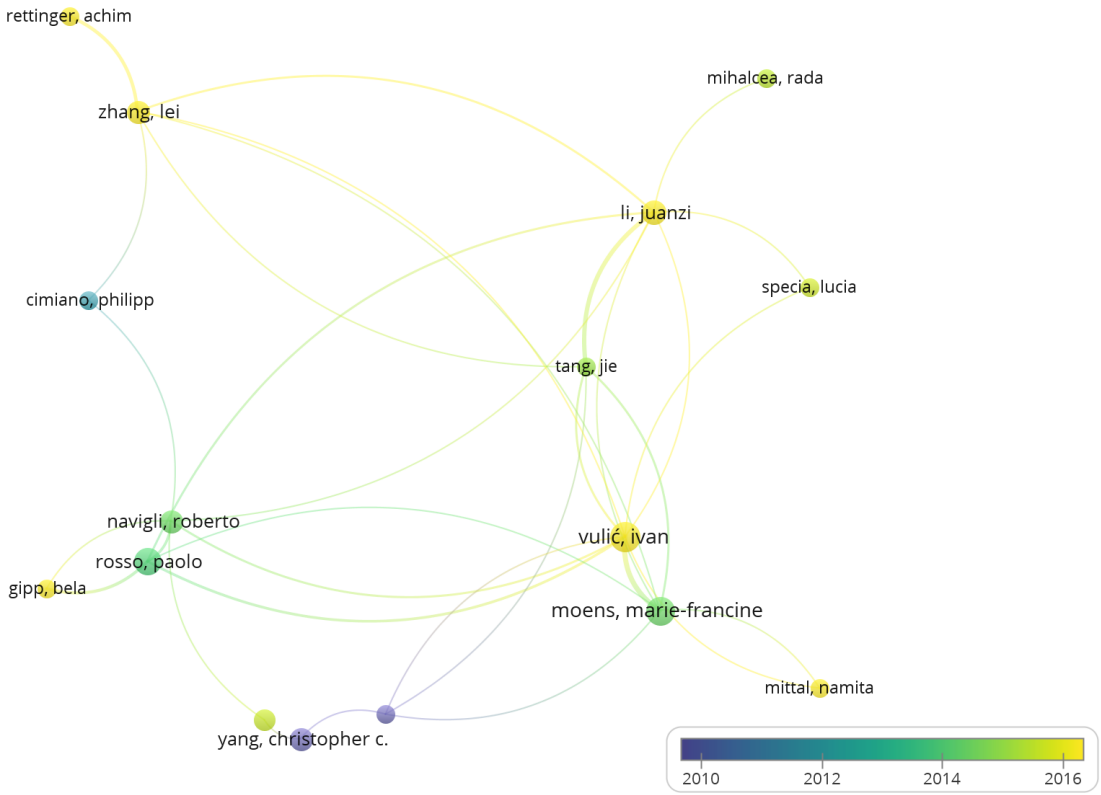

Fig. 7. Overlay visualization of the most cited authors

The geographical aspects of citation were considered. A citation network for countries was built with a minimum of five publications. 34 countries were identified (Fig. 8). It is seen that the geography of research is expanding. So, in the last goals, Brazil, Czech Republic, Iran, Egypt, Tunisia have joined the research.

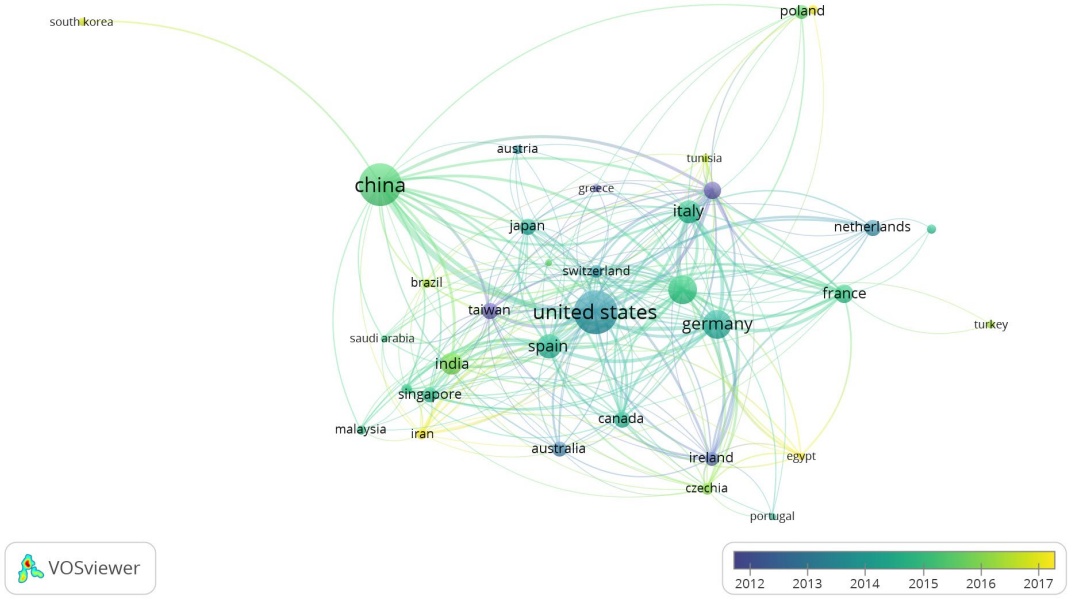

Fig. 8. Overlay visualization of citation by affiliation of authors

5. Conclusions

A bibliometric study of publications on the topic "Cross-lingual Semantic Similarity", available in the Dimensions database, was carried out. In recent years, there has been a significant increase in research. Visualization of scientific networks using VOSviewer has shown fragmentation of research, small research groups have been identified.

The visualization of a network of co-authorship across organizations showed limited university interaction on cross-language semantic similarities. The largest cluster

included 11 European universities and organizations from Ireland, Italy, Germany, Spain, Scotland, and Great Britain.

Visualization of the co-authorship network by country showed that 35 countries interact in research, countries are connected in five clusters. The two largest clusters included 9 countries. In the largest clusters, including 9 countries, the leading ones were the USA and China, Great Britain and Spain.

The visualization of the citation network revealed 298 of the most cited authors out of 2050. The most cited author is Navigli, Roberto (759 citations). More than 200 citations from Rosso, Paolo (239) and Moens, Marie-Francine (216) [2].

Overlay visualization made it possible to evaluate the citation trends of the authors; it turned out that the most cited author, Navigli, Roberto, is also the founder of research in this field [29, 30].

The most recent cited authors are Mittal Namitali (2017 citation), Rettinger Achim, Gipp Bela, Li Juanzi and Zhan Lei (2016 citation).

Consideration of the geographical aspects of citation showed an expansion of the geography of research. So, in the last goals, Brazil, Czech Republic, Iran, Egypt, Tunisia have joined the research.

Acknowledgment

The reported study was funded by RFBR according to the research projects № 18-07-00225, 18-07-00909, 18-07-01111 and 20-04-60185.

References

[1] Rajat Pandit, R., Sengupta, S., Naskar, S.K., Dash, N.S. and Sardar, M.M. (2019). Improving Semantic Similarity with Cross-Lingual Resources: A Study in Bangla - A Low Resourced Language. Informatics, 6, 19; doi:10.3390/informatics6020019

[2] Vulic, I., De Smet, W., and Moens, M.-F. (2011). Identifying word translations from comparable corpora using latent topic models. In Proceedings of ACL, pages 479-484.

[3] Prochasson, E. and Fung, P. (2011). Rare word translation extraction from aligned comparable documents. In Proceedings of ACL, pages 1327-1335.

[4] Hotho, A., Nürnberger, A. and Paaß, G. (2005). A brief survey of text mining. In Ldv Forum, Vol. 20(1), p. 19-62.

[5] Hassani, H., Beneki, C., Unger, S., Mazinani, M.T. and Yeganegi, M.R. (2020). Text Mining in Big Data Analytics. Big Data Cogn. Comput. 2020, 4, 1; doi:10.3390/bdcc4010001.

[6] Porter, A. L. (2005). Tech Mining. Competitive Intelligence Magazine. 8 (1): 30-37.

[7] Ali, A., Alfayez, F. and Alquhayz, H. (2018). Semantic Similarity Measures Between Words: A Brief Survey. Sci. Int. (Lahore),30(6), 907-914, 2018.

[8] Wang, H. C., Chi, Y. C. and Hsin, P. L. (2018). Constructing Patent Maps Using Text Mining to Sustainably Detect Potential Technological Opportunities. Sustainability, 10, 3729; doi:10.3390/su10103729.

[9] Grappiolo, C., van Gerwen, E., Verhoosel, J. and Somers, L. (2019). The Semantic Snake Charmer Search Engine: A Tool to Facilitate Data Science in High-tech Industry Domains. In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR '19). Association for Computing Machinery, New York, NY, USA, 355-359. DOI:https://doi.org/10.1145/3295750.3298915.

[10] Jarmasz, M. and Szpakowicz, S. (2003). Roget's Thesaurus and Semantic Similarity. Recent Adv. Nat. Lang. Process. III Sel. Pap. from RANLP , vol. 111, 2004.

[11] Islam, A. and Inkpen, D. (2012). Unsupervised Near-Synonym Choice using the Google Web 1T. ACM Trans. Knowl. Discov. Data, vol. V, no. June, pp. 1-19.

[12] O'Shea, J., Bandar, Z., Crockett, K., and McLean, D. (2008). A Comparative Study of Two Short Text Semantic Similarity Measures. In Agent and Multi-Agent Systems: Technologies and Applications, vol. 4953, N. Nguyen, G. Jo, R. Howlett, and L. Jain, Eds. Springer Berlin Heidelberg, pp. 172-181.

[13] Li, H. and Xu, J. (2014). Semantic matching in search. Foundations and Trends in Information Retrieval, 7(5):343-469.

[14] Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. Cognitive science, 34(8), 1388-1429.

[15] Chen, B. (2009). Latent topic modelling of word co-occurence information for spoken document retrieval. In IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2009, no. 2, pp. 3961-3964.

[16] Kenter, T., Rijke, M. de (2015). Short Text Similarity with Word Embeddings. CIKM '15 Proceedings of the 24th ACM International on Conference on Information and Knowledge Management October 19-23, Melbourne, Australia. Pp. 1411-1420.

[17] Atoum, I. (2016). Efficient Hybrid Semantic Text Similarity using Wordnet and a Corpus. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 9, pp.124-130.

[18] Magerman, T., Van Looy, B., Baesens, B. and Debackere, K. (2011). Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents. Department Of Managerial Economics, Strategy And Innovation (MSI), October, 77 p.

[19] Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. (2014). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. SemEval-2014.

[20] Batet, M. and Sánchez, D. (2015). Ontology Selection for Semantic Similarity Assessment. ICAART 2015, At Lisbon, Portugal, Volume: 2 https://www.researchgate.net/publication/283877653

[21] Liu, H., Wang, P. (2014). Assessing Text Semantic Similarity Using Ontology. Journal Of Software, vol. 9, no. 2, pp.490-497.
[22] Maheswari, J.U., Karpagam, G.R., Indhumathy, S. (2014). Comparison of Web Service Similarity-Assessment Methods. International Journal of Computer Applications (0975 - 8887) Volume 98 - No.22.
[23] Moen, H. (2016). Distributional Semantic Models for Clinical Text Applied to Health Record Summarization Thesis for the Degree of Philosophiae Doctor Trondheim, May NTNU (Norwegian University of Science and Technology Faculty of Information Technology), 93 p.
[24] Guessoum, D., Miraoui, M., Tadj, C. (2015). Survey Of Semantic Similarity Measures In Pervasive Computing. International Journal On Smart Sensing And Intelligent Systems Vol. 8, no. 1, pp.125-158.
[25] Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In Proceedings of ICLR 2017. https://openreview.net/pdf?id=SyK00v5xx.
[26] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. CoRR abs/1705.02364. http://arxiv.org/abs/1705.02364.
[27] Pagliardini, M., Gupta, P., and Jaggi, M. (2017). Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. arXiv https://arxiv.org/pdf/1703.02507.pdf.
[28] Ferrero, J., Besacier, L., Schwab, D., and Agnes, F. (2017). Using Word Embedding for Cross-Language Plagiarism Detection. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, (EACL 2017). Association for Computational Linguistics, Valencia, Spain, volume 2, pages 415-421. http://aclweb.org/anthology/E/E17/E17-2066.pdf.
[29] Camacho-Collados, J. and Navigli, R. (2016). Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP. Berlin, Germany, pages 43-50.
[30] Camacho-Collados, J., Taher Pilehvar, M., Collier, N., and Navigli, R. (2017). SemEval-2017 Task 2: Multilingual and cross-lingual semantic word similarity. In Proceedings of SemEval. Vancouver, Canada.
[31] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
[32] Van Eck, N.J., and Waltman, L. How to normalize cooccurrence data? An analysis of some well-known similarity measures. 2009. Journal of the American Society for Information Science and Technology, 60(8), 1635-1651.

About the autors

Khakimova Aida Kh., PhD, docent, Kama Institute (Naberezhnye Chelny, Russia), ANO «Scientific and Research Center for Information in Physics and Technique» (Nizhny Novgorod, Russia), E-mail: aida_khatif@mail.ru